Empirical evidence of an association between internal validity and effect size in randomized controlled trials of low-back pain

van Tulder, M.W.; Suttorp, M.; Morton, S.; Bouter, L.M.; Shekelle, P.

published in
Spine
2009

DOI (link to publisher)
10.1097/BRS.0b013e3181ab6a78

document version
Publisher's PDF, also known as Version of record

SPINE Volume 34, Number 16, pp 1685–1692

Empirical Evidence of an Association Between Internal Validity and Effect Size in Randomized Controlled Trials of Low-Back Pain

Maurits W. van Tulder, PhD,*† Marika Suttorp, MSc,‡ Sally Morton, PhD,§ Lex M. Bouter, PhD,† and Paul Shekelle, MD, PhD‡¶

Study Design. We conducted a methodologic study.

Objective. The objective of this study was to assess the validity of the criteria list recommended by the Cochrane Back Review Group Editorial Board by evaluating whether individual items and a total score are associated with effect sizes in randomized controlled trials of back-pain interventions.

Summary of Background Data. There is concern that studies of low methodologic quality may exaggerate the effectiveness of treatments for low back pain. We performed this study to examine the association between a common measure of internal validity and the reported magnitude of treatment effects.

Methods. We assessed the relationship between the 11 items contained in the Cochrane Back Review Group Internal Validity checklist and effect size in randomized trials of interventions for back pain. Of 267 trials in 15 Cochrane reviews that were eligible for inclusion, 51 were excluded, leaving 216 trials included in the analysis. The scores on the 11 items for each trial were taken from the original review. We extracted effect sizes from each low back pain trial.

Results. We found that trials that fulfilled a specific item had smaller effect sizes compared with trials that did not fulfill that item for 10 of the 11 items, and for 6 of the criteria, the absolute difference in effect sizes was 0.10 or greater. The 95% confidence interval of the difference in effect sizes crossed the null value in each case. The number of items fulfilled showed that trials with higher scores consistently reported smaller effect sizes than trials with lower scores. At the thresholds of 5 or 6 items fulfilled, the difference in effect sizes was 0.20 in each case (95% confidence intervals 0.05–0.35 and 0.06–0.34, respectively). Stratified analyses did not support confounding by intervention.

Conclusion. We conclude that the 11-item Internal Validity Checklist is associated with effect size in randomized trials of interventions for back pain, and that our data support the use of a sum score of the number of fulfilled items in this list.

Key words: Cochrane collaboration, randomized trials, low back pain, bias, effect size. **Spine 2009;34:1685–1692**

There is concern that studies of low methodologic quality may exaggerate the effectiveness of treatments for low back pain. We performed this study to examine the association between a common measure of internal validity and the reported magnitude of treatment effects. The measurement of the internal validity of randomized controlled trials (RCTs) is a key component of assessing the evidence about health care interventions. How to measure internal validity remains controversial. Some studies have reported empirical evidence relating to bias in RCTs with some individual criteria such as concealment of treatment allocation and blinding of patients,[1,2] but others have reported no significant association between individual criteria and estimates of treatment effect.[3] Further compounding the problem is the difficulty in summarizing the assessment of the internal validity of an RCT when the trial varies in its compliance with individual criteria. Jadad et al[4] proposed and psychometrically evaluated a scale to distinguish between trials with "high" and "low" internal validity, which used 3 items: randomization, double-blinding, and a description of dropouts and withdrawals. Moher *et al* reported that there was evidence to support the relationship between risk of bias and effect size,[5] and the Jadad scale has become widely used as a summary measure of quality of RCTs included in systematic reviews and meta-analyses. However, as this scale was developed to assess drug trials and double-blinding accounts for 40% of the score, the Jadad scale is less useful for interventions where double-blinding may be difficult. Furthermore, there is evidence of a relationship with quality for criteria not included in this scale (such as concealment of random allocation).[1,2]

Almost simultaneously with the development of the Jadad scale, Verhagen *et al* developed a list of 9 items focused specifically on internal validity. They used a formal Delphi process of 3 rounds, with input from leading experts around the world.[6] Jadad *et al* and Verhagen *et al*

From the *Department of Health Sciences, EMGO Institute for Health and Care Research, VU University, Amsterdam, The Netherlands; †Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands; ‡RAND Health, Santa Monica, CA; §RTI International, Research Triangle Park, NC; and ¶West Los Angeles VA Medical Center, Los Angeles, CA.

The manuscript submitted does not contain information about medical device(s)/drug(s).

No funds were received in support of this work. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.

Maurits van Tulder, Lex Bouter, and Paul Shekelle were responsible for conception and design of the study; Maurits van Tulder, Marika Suttorp, and Paul Shekelle were involved in acquisition of data; Marika Suttorp and Sally Morton had full access to all of the data in the study, conducted the analysis and take responsibility for the integrity of the data and the accuracy of the data analysis; all authors were involved in interpretation of data; Maurits van Tulder and Paul Shekelle drafted the article; all authors were involved in revising the article critically for important intellectual content and in final approval of the version to be published.

Address correspondence and reprint requests to Maurits van Tulder, PhD, Department of Health Sciences, VU University Amsterdam, The Netherlands; E-mail: maurits.van.tulder@falw.vu.nl

developed their criteria lists in fields other than back pain and it was unclear whether these lists were also valid for back-pain studies.

Koes *et al* published the first systematic reviews in the field of back pain in 1991.[7,8] The assessment of the methodologic quality of RCTs included in these reviews was done using a list of criteria that was based on generally accepted principles of intervention research, from textbooks on clinical epidemiology. This list included items related to internal validity, external validity, and precision.

Since that time, the potential hazards of using scales were reported by Juni *et al*, who applied 25 different published scales to 17 trials of low molecular weight heparin. Depending on which scale was used to classify trials as having high or low internal validity, the conclusions of the review could be completely reversed. [9] Juni *et al* postulated that their results were due in part to the inclusion in most scales of items unrelated to internal validity, such as items about ethics or external validity.

In 1997, the Cochrane Back Review Group (CBRG) Editorial Board published method guidelines for systematic reviews in the field of spinal disorders, which they updated in 2003.[10] They recommended that 11 criteria be used as the standard measure to assess the methodologic quality, or internal validity, of RCTs.[11] This list was based on the criteria in Jadad *et al's*[4] and Verhagen *et al's* lists[6] and was a modification of the list used by Koes *et al*.[7,8] These criteria were used in 77% of the CBRG reviews published in The Cochrane Library 2005, issue 3. Furthermore, the CBRG Editorial Board recommended using a score as a measure of overall internal validity.[8] Compliance with 6 criteria, resulting in a score of 6, (similar to the threshold used in the Jadad scale of about half the items being successfully met) was suggested as a means of distinguishing high from low quality trials.

The objective of this study was to assess the validity of the criteria list recommended by the CBRG Editorial Board by evaluating whether individual items and a total score are associated with effect sizes in RCTs of back-pain interventions.

■ Materials and Methods

Trial Selection

Eligible Reviews. All CBRG reviews of a nonsurgical treatment for nonspecific low back pain that were present in the Cochrane Library 2005, issue 3, were eligible for inclusion.

Eligible Trials. All RCTs included in the selected reviews were eligible if they included pain, function, or similar improvement measure as an outcome. Comparisons could be between a treatment and placebo, usual care, "no treatment," or another treatment. Trials were excluded if they did not present data in such a way that an effect size could be calculated.

Quality Assessment

The list of criteria recommended by the CBRG Editorial Board and the definition of the items are presented in Tables 1 and 2.

Table 1. Cochrane Back Review Group Internal Validity Checklist[11]

A	Was the method of randomization adequate?
B	Was the treatment allocation concealed?
C	Were the groups similar at baseline regarding the most important prognostic indicators?
D	Was the patient blinded to the intervention?
E	Was the care provider blinded to the intervention?
F	Was the outcome assessor blinded to the intervention?
G	Were co-interventions avoided or similar?
H	Was the compliance acceptable in all groups?
I	Was the drop-out rate described and acceptable?
J	Was the timing of the outcome assessment in all groups similar?
K	Did the analysis include an intention-to-treat analysis?

The quality assessment scores reported in the original Cochrane reviews were used. All items were scored "yes" if clearly fulfilled, "no" if clearly not fulfilled or "don't know" if it was unclear from the paper if the item was fulfilled or not. Quality assessment in all Cochrane reviews was done independently by 2 review authors and a consensus meeting was used to resolve any disagreement between them. The first author (MvT) of the

Table 2. Operationalization of the Criteria List

A	A random (unpredictable) assignment sequence. Examples of adequate methods are computer generated random No. table and use of sealed opaque envelopes. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should not be regarded as appropriate
B	Assignment generated by an independent person not responsible for determining the eligibility of the patients. This person has no information about the persons included in the trial and has no influence on the assignment sequence or on the decision about eligibility of the patient
C	In order to receive a yes, groups have to be similar at baseline regarding demographic factors, duration and severity of complaints, percentage of patients with neurologic symptoms, and value of main outcome measure(s)
D	The reviewer determines if enough information about the blinding is given in order to score a yes
E	The reviewer determines if enough information about the blinding is given in order to score a yes
F	The reviewer determines if enough information about the blinding is given in order to score a yes
G	Co-interventions should either be avoided in the trial design or similar between the index and control groups
H	The reviewer determines if the compliance to the interventions is acceptable, based on the reported intensity, duration, no. and frequency of sessions for both the index intervention and control intervention(s)
I	The No. participants who were included in the study but did not complete the observation period or were not included in the analysis must be described and reasons given. If the percentage of withdrawals and drop-outs does not exceed 20% for short-term follow-up and 30% for long-term follow-up and does not lead to substantial bias a yes is scored. (N.B. these percentages are arbitrary, not supported by literature)
J	Timing of outcome assessment should be identical for all intervention groups and for all important outcome assessments
K	All randomized patients are reported/analyzed in the group they were allocated to by randomization for the most important moments of effect measurement (minus missing values) irrespective of noncompliance and co-interventions

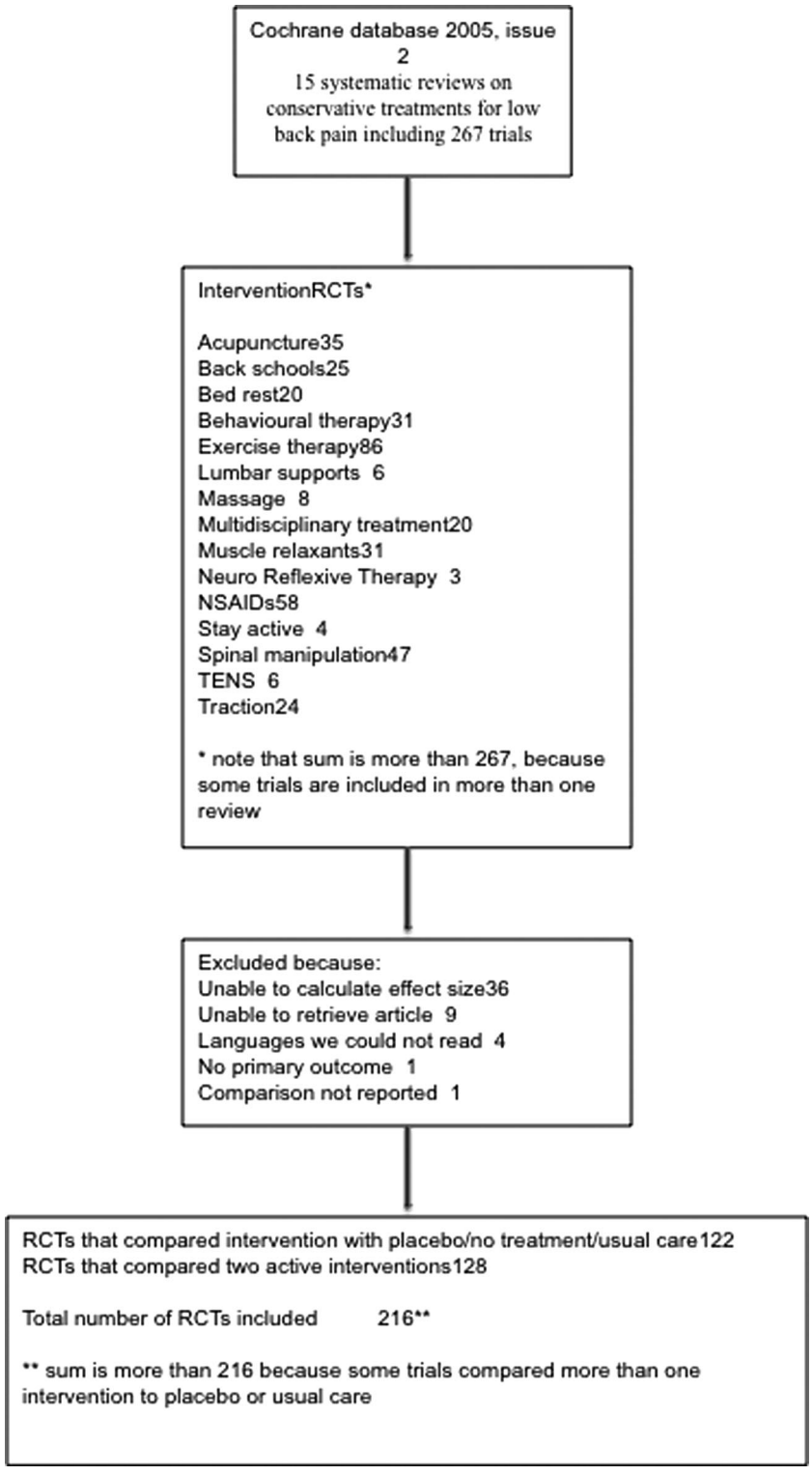

Figure 1. Flow chart of trial selection.

present article was involved in the quality assessment of 169 of the 216 trials (78%); the last author (P.S.) in the assessment of 33 (15%).

Thirty-nine (18%) trials were included in more than one Cochrane review. Five of these trials disagreed on the quality score for one or more items.[12–16] For these trials, the highest quality rating was used.

Effect Sizes

Statistical data were collected for 1 outcome per trial. Means and standard deviations were collected for continuous outcomes (pain and function) and proportions for dichotomous outcomes (*e.g.*, proportion who improved). In cases where more than 1 outcome was reported, we selected one using a hierarchy: (1) short-term pain, (2) short-term function, (3) long-term pain, (4) long-term function, (5) number of patients improved in the short-term, (6) number of patients improved in the long-term. Short-term was defined as closest to 6 weeks and long-term as closest to 1 year. If a trial did not report any of these outcomes, we did not include it. For trials that reported a mean outcome but no standard deviation, we estimated the

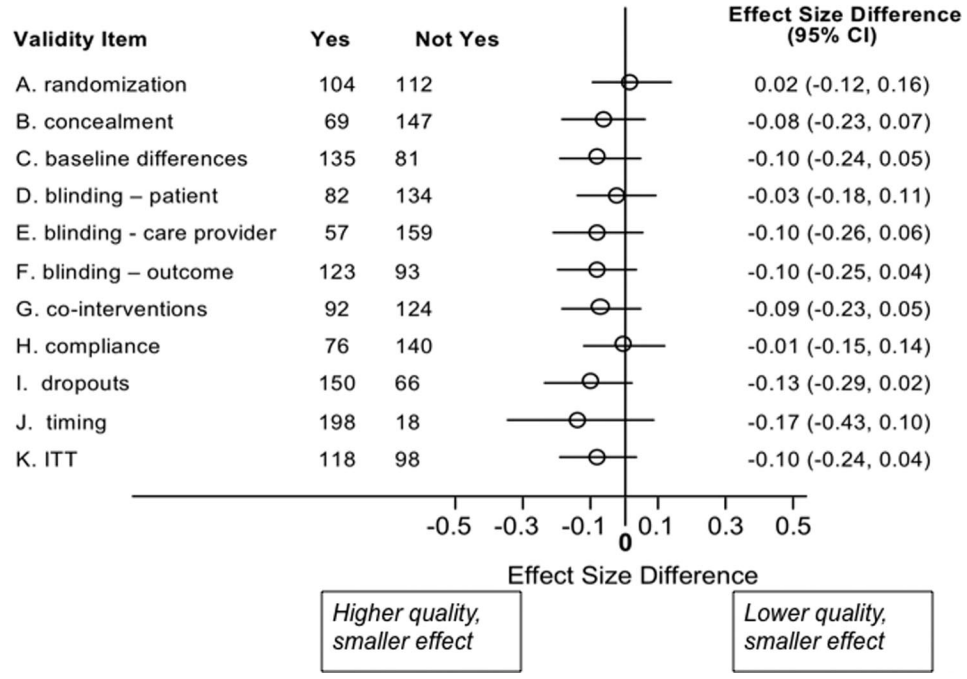

Figure 2. Difference in effect sizes on studies based on quality criteria.

standard deviation by taking the mean standard deviation, weighted by the relevant treatment group's sample size, across all other trials that reported standard deviations for that outcome.[17]

Effect sizes were assessed because they convert different scales into the same metric, so that they can be compared. Effect sizes for continuous outcomes were calculated by dividing the difference between the follow-up means of the 2 intervention groups by the pooled follow-up standard deviations of the 2 intervention groups.[18] Effect sizes for dichotomous outcomes were calculated by using the Kraemer and Andrews estimator,[19] which uses the inverse normal cumulative distribution function to transform from a dichotomous to a continuous scale. We performed analyses using both the actual (either positive or negative) and the absolute (negative transformed to positive) value of the effect size. The difference between results using either effect size was small, since relatively few trials reported negative effect sizes (only 15% of placebo controlled trials). Results reported are for the absolute value of the effect size.

We included all trials in the analysis. We also ran each analysis separately for the intervention *versus* placebo/usual care trials and the intervention *versus* intervention trials. In the case of a multi-intervention trial, we only used 2 intervention arms, or 1 comparison, in the analysis to avoid double counting. We selected the comparison with the largest absolute effect size.

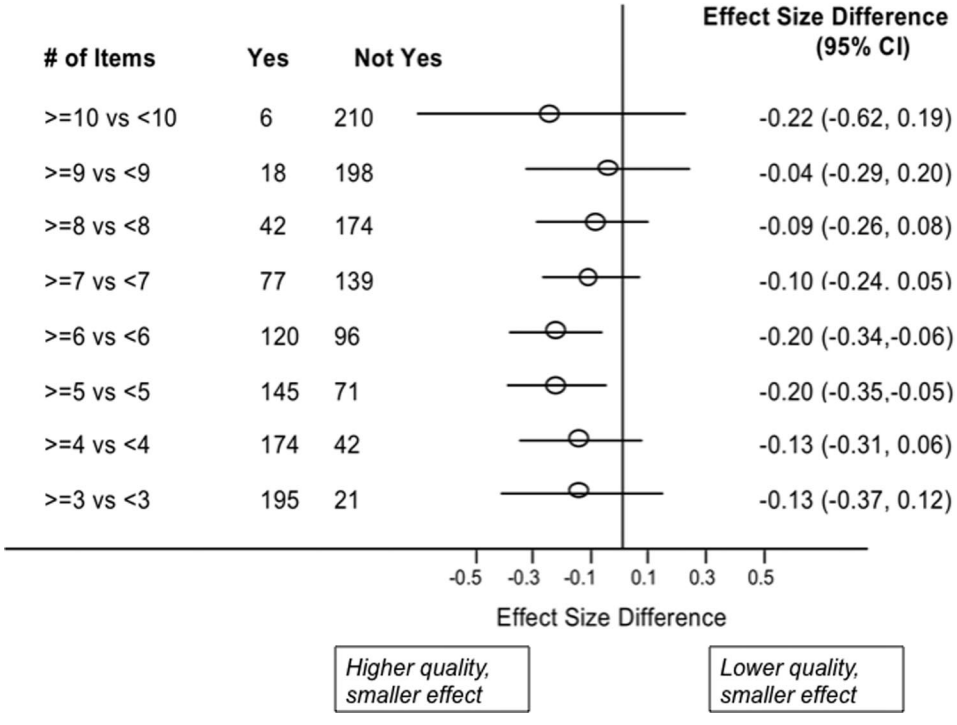

Figure 3. Difference in effect sizes for studies meeting or exceeding differing sum scores of the criteria scored Yes.

Table 3. Meta-Regression Analysis of Sum Score Association With Differences in Effect Sizes (DES) By Intervention

	Intervention															
	Acupuncture				Back School				Behavior Therapy				Exercise Therapy			
Sum Score	Yes	No	DES*	95% CI	Yes	No	DES*	95% CI	Yes	No	DES*	95% CI	Yes	No	DES*	95% CI
≥2 *vs.* <2	24	1	−0.88	(−1.98, 0.23)	18	1	0.38	(−0.72, 0.47)	15	1	0.33	(−0.67, 1.32)	46	2	−0.52	(−1.28, 0.24)
≥3 *vs.* <3	20	5	−0.40	(−0.89, 0.09)	15	4	−0.20	(−0.76, 0.36)	13	3	0.18	(−0.49, 0.85)	43	5	−0.44	(−0.97, 0.10)
≥4 *vs.* <4	17	8	−0.29	(−0.73, 0.15)	10	9	−0.04	(−0.48, 0.40)	12	4	−0.12	(−0.76, 0.52)	38	10	−0.11	(−0.50, 0.28)
≥5 *vs.* <5	15	10	−0.29	(−0.69, 0.10)	8	11	−0.12	(−0.54, 0.31)	10	6	−0.33	(−0.89, 0.23)	34	14	−0.41	(−0.74, −0.09)
≥6 *vs.* <6	13	12	−0.20	(−0.57, 0.17)	5	14	0.02	(−0.42, 0.46)	5	11	−0.15	(−0.70, 0.40)	28	20	−0.64	(−0.92, −0.36)
≥7 *vs.* <7	10	15	−0.13	(−0.52, 0.26)	3	16	−0.04	(−0.60, 0.53)	2	14	0.27	(−0.55, 1.09)	20	28	−0.54	(−0.82, −0.27)
≥8*vs.* <8	7	18	−0.19	(−0.63, 0.24)	2	17	−0.01	(−0.71, 0.69)	2	14	0.27	(−0.57, 1.11)	10	38	−0.33	(−0.66, 0.00)
≥9 *vs.* <9	2	23	0.08	(−0.66, 0.83)	0	19	—	—	1	15	−0.15	(−1.21, 0.91)	6	42	−0.32	(−0.73, 0.09)
≥10 *vs.* <10	0	25	—	—	0	19	—	—	1	15	−0.15	(−1.21, 0.91)	2	46	−0.30	(−0.96, 0.35)
11 vs. <11	0	25	—	—	0	19	—	—	0	16	—	—	1	47	−0.52	(−1.42, 0.38)

*Negative values mean studies scoring higher (lower risk of bias) have lower estimates of effect than studies scoring lower (higher risk of bias).

Statistical Analysis

The difference of effect sizes between studies with the quality item scored yes and those with the quality item scored "not yes" (no or don't know) was used as a measure of bias associated with that quality item. The difference was estimated using meta-regression.[20] A meta-regression was conducted separately for each quality item. The coefficient from each regression estimates the difference in effect sizes between studies with the quality item scored yes *versus* those in which it is scored not yes. A negative coefficient indicates that studies with the quality item scored yes have smaller effect sizes than those that scored not yes. All analyses were conducted in Stata 9.2.[21]

A sum score was calculated by adding the quality items that were scored yes. The difference in effect sizes of studies above and below all possible thresholds of 3 or more positive scores (*i.e.*, 4 or more *vs.* less than 4; 5 or more *vs.* less than 5; *etc.*) was compared using the methods outlined above. We also performed our analyses by intervention type. Each intervention type that had 10 or more trials that could be assessed was included in this analysis. We conducted a meta-regression for each threshold score. Each meta-regression was a fully-interacted model, containing a main effect for having the threshold or higher number of quality items *versus* not, main effects for each intervention type, and the interactions between the threshold and each intervention. We report estimates for those interventions that had at least 1 study that met the threshold and 1 study that did not.

■ Results

Of the 15 Cochrane reviews[22–36] on nonsurgical treatment for nonspecific low back pain, 267 trials were eligible for inclusion (Figure 1). Of these, 36 were excluded because they did not provide sufficient data to calculate an effect size. Four articles were in languages we could not read, 1 did not have a primary outcome and 1 did not report on a comparison of interest. We were unable to obtain 9 articles. Thus, 216 trials were included in our analysis: 122 where the back pain treatment was compared to placebo or usual care and 128 trials where the treatment was compared to another back-pain treatment (which add up to more than 216 because some trials compared more than 1 intervention to placebo or usual care). Of these 216 trials, 139 (64%) reported short-term pain outcomes, 14 (6%) reported short-term function outcomes, 12 (6%) reported long-term pain outcomes, 7 (3%) report long-term function outcomes, 42 (19%) reported short-term improvement, and 2 (1%) reported long-term improvement. Fifty-four trials reported data on more than 1 comparison. The number of internal validity quality criteria fulfilled ranged from 1 to 11, with a mean of 5.6.

Figure 2 presents the difference in effect sizes between studies scoring yes or not yes on each of the 11 quality criteria. There were smaller effect sizes for studies scoring yes compared to studies scoring not yes for 10 of the 11 criteria, and for 6 of the criteria, the absolute difference in effect sizes was 0.10 or greater. The 95% confidence interval of the difference in effect sizes crossed the null value in each case.

Figure 3 presents the difference in effect sizes for studies meeting or exceeding a particular score (threshold). For each threshold, studies scoring above the threshold had smaller effect sizes than studies scoring below the threshold. The difference in effect sizes between trials with higher and lower scores is 0.20 for thresholds of 5 and 6 and the 95% confidence intervals of these thresholds do not cross the null value.

Table 3 presents the results of the meta-regression analysis. It includes all interventions for which there were at least 10 trials (except bed rest, which was frequently the comparison group for trials of an active therapy; and multidisciplinary treatment, which was too heterogeneous to classify as a therapy), and all thresholds. As expected, there is variability in the difference in effect sizes with the smaller sample sizes in each comparison. However, this variability does not reflect chance. For 38 of the 56 comparisons where there was at least 1 study scoring above and below the threshold, higher quality studies had a smaller estimated effect size than lower quality studies, compared to 17 times where the opposite was observed (sign test, $P = 0.006$). In comparisons where there were at least 25% of the studies scoring above or below the threshold, 19 of 21 higher quality

Table 3. Continued

Intervention											
Manipulation				Muscle Relaxation				NSAIDs			
Yes	No	DES*	95% CI	Yes	No	DES*	95% CI	Yes	No	DES*	95% CI
21	1	0.4	(−0.70, 1.51)	25	0	—	—	36	1	0.24	(−0.80, 1.27)
20	2	0.37	(−0.39, 1.14)	25	0	—	—	36	1	0.24	(−0.80, 1.27)
19	3	−0.39	(−1.11, 0.33)	24	1	0.30	(−0.66, 1.25)	34	3	0.12	(−0.45, 0.69)
16	6	−0.42	(−0.94, 0.11)	23	2	0.10	(−0.62, 0.83)	26	11	−0.19	(−0.52, 0.14)
13	9	−0.51	(−0.93, −0.09)	20	5	0.04	(−0.42, 0.49)	24	13	−0.15	(−0.46, 0.16)
9	13	−0.23	(−0.64, 0.18)	11	14	0.18	(−0.21, 0.57)	13	24	−0.02	(−0.34, 0.29)
5	17	−0.12	(−0.58, 0.35)	4	21	−0.34	(−0.88, 0.20)	7	30	0.14	(−0.27, 0.54)
1	21	−0.46	(−1.37, 0.44)	0	25	—	—	3	34	−0.02	(−0.60, 0.55)
0	22	—	—	0	25	—	—	2	35	0.03	(−0.66, 0.72)
0	22	—	—	0	25	—	—	0	37	—	—

studies had estimated effect sizes that were smaller than lower quality studies (sign test, $P = 0.0002$). These data support the conclusion that a summary score of these 11 items is associated with bias in effect size, regardless of the intervention studied or threshold chosen.

■ Discussion

The 2 principal findings of our study are: that reports of RCTs of low back pain treatments consistently report smaller effect sizes if they fulfill most of the individual items in the CBRG internal validity checklist; and at a threshold of 5 or 6 fulfilled criteria, trials fulfilling more criteria (low risk of bias) have effect sizes that are up to 0.20 lower than trials meeting fewer criteria (high risk of bias). When 6 criteria were fulfilled, the average effect size of trials meeting fewer than 6 criteria was 0.58, whereas the average effect size of trials fulfilling more than 6 criteria was 0.38, meaning that trials with a higher risk of bias reported effect sizes that were, on average, 50% greater than estimates reported from trials with a lower risk of bias. Our results are not substantially affected by the contrast studied or the treatment assessed.

Prior studies of the effects of individual internal validity items are mixed. Juni *et al* reported a meta-analysis of 3 items— generation of the allocation sequence, concealment of treatment allocation, and double-blinding—that were assessed in 4 methodologic studies that gauged their relative importance in a large number of clinical trials, while avoiding confounding by disease or by intervention.[37] In the original methodologic studies, these items were mostly associated with smaller treatment effects (but not entirely so—in one methodologic study double-blinding was associated with larger treatment effects[5]), but in about half of the instances, these differences were not statistically significant. In their pooled analysis of the ratio of odds ratios, the generation of allocation sequence had a pooled value of 0.81 (95% CI: 0.60–1.09), concealment of treatment allocation had a pooled value of 0.70 (95% CI: 0.62–0.80), and double-blinding had a pooled value of 0.86 (95% CI: 0.74–0.99). Balk *et al* examined 24 quality criteria, assessing 276 trials in 4 clinical conditions, and found no criterion that was associated with a statistically significantly difference in treatment effects.[3] Eight of these 24 criteria were identical or related to criteria in the CBRG internal validity checklist. Our findings for individual items are similar; none were associated with a statistically significantly difference in treatment effect, although effect sizes were consistently smaller if items were fulfilled.

Our results for the summary score are novel. Prior attempts at assessing the effect of a summary scale yielded no association between a scale developed by Chalmers and coworkers,[38] and a significant association between higher scores on the Jadad scale and lower estimates of treatment effects.[5] Juni *et al* demonstrated in one meta-analysis that the application of 25 different scales yielded conflicting results.[10] These data on variability contribute to many authorities' lack of enthusiasm for scales, who advise instead an assessment of trials according to individual quality components. However, the problem remains of how to assess the validity of the results of different trials when they vary on multiple individual quality criteria. This is particularly true in narrative reviews of trials, where there is no meta-analytic pooled estimate of effect to provide the basis for multiple sensitivity analyses of the effect of individual components on treatment estimates. Even with the use of only 3 quality criteria, there are 8 different potential combinations of "yes" and "no" scores, making it a challenge to assess the trade-offs in internal validity between trials with different combinations of yes and no scores. Without an empirically-validated summary measure of internal validity, such assessments will be done on an ad hoc basis, using decision-making rules that are idiosyncratic and possibly unstated, or not done at all. The past problems with scales have been attributed to the heterogeneous nature of their criteria, with some scales including criteria related to external validity, interpretation, or ethical issues, and that many scales include items for which there is little evidence that they are related to the internal validity of the trial.[10,37] It is possible that the more favorable results we found are because we restricted items

related only to internal validity, and that the items had empirical evidence of content validity from their selection as part of the Delphi list process.

Our study has 2 primary limitations. The first is that we relied on the internal validity assessment scores of the authors of the original CBRG review. As such, we have no measure of the inter-rater reliability of the scoring of individual items across reviews. However, Verhagen *et al* showed that the inter-rater reliability of 20 reviewers using the criteria list was high,[39] and Balk *et al* in their study of similar criteria also reported high inter-rater agreement. Furthermore, the first author of the present article was involved in the quality assessment of 78% of the trials. The second limitation is that all of the clinical trials assessed patients with low back pain. Previous assessments of the relationship between quality criteria or scales and the estimate of treatment effects have been critiqued if they were subject to potential confounding by condition or by intervention. Our study included trials of many different interventions, and our meta-regression analyses support the validity of our findings for interventions as disparate as pharmaceuticals, behavioral therapies, and complementary and alternative medicine therapies. However, we cannot fully exclude the possibility of confounding by condition, although we do note that the items in the CBRG internal validity list seem broadly applicable and not specific to back pain. These items have indeed been used in systematic reviews of other conditions.

In summary, we found evidence that a small number of items in the CBRG internal validity checklist that are fulfilled appears to be associated with biased treatment effects. Studies fulfilling fewer than 5 or 6 of 11 items had significantly higher estimates of treatment effects. Our data support the use of a score, reached by adding the number of items in this list that are fulfilled, to indicate the internal validity of randomized trials of treatments for patients with low back pain. We believe that it's likely that the CBRG internal validity list is also useful for the assessment of trials in other conditions.

■ Key Points

- There is concern that studies of low methodologic quality may exaggerate the effectiveness of treatments for low back pain.
- We assessed the relationship between the 11 items contained in the Cochrane Back Review Group Internal Validity checklist and effect size in randomized trials of interventions for back pain.
- Two hundred sixteen trials of 15 Cochrane reviews were included in the analysis.
- The number of items fulfilled showed that trials with higher scores consistently reported smaller effect sizes than trials with lower scores. At the thresholds of 5 or 6 items fulfilled, the difference in effect sizes was 0.20 in each case.
- The 11-item Internal Validity Checklist is associated with effect size in randomized trials of interventions for back pain, and that our data support the use of a sum score.

References

1. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy, I: medical. *Stat Med* 1989;8:441–54.
2. Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–12.
3. Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973–82.
4. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996; 17:1–12.
5. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609–13.
6. Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol* 1998;51:1235–41.
7. Koes BW, Assendelft WJ, van der Heijden GJ, et al. Spinal manipulation and mobilisation for back and neck pain: a blinded review. *BMJ* 1991;303: 1298–303.
8. Koes BW, Bouter LM, Beckerman H, et al. Physiotherapy exercises and back pain: a blinded review. *BMJ* 1991;302:1572–6.
9. Juni P, Witschi A, Bloch R, et al. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054–60.
10. van Tulder MW, Assendelft WJ, Koes BW, et al; the Editorial Board of the Cochrane Back Review Group. Method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group for Spinal Disorders. *Spine* 1997;22:2323–30.
11. van Tulder M, Furlan A, Bombardier C, et al. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. *Spine* 2003;28:1290–9.
12. Alaranta H, Rytökoski U, Rissanen A, et al. Intensive physical and psychosocial training program for patients with chronic low back pain: a controlled clinical trial. *Spine* 1994;19:1339–49.
13. Bendix AF, Bendix T, Lund C, et al. Comparison of three intensive programs for chronic low back pain patients: a prospective, randomized, observer-blinded study with one-year follow-up. *Scand J Rehabil Med* 1997;29:81–9.
14. Berry H, Hutchinson DR. Tizanidine and ibuprofen in acute low-back pain: results of a double-blind multicentre study in general practice. *J Int Med Res* 1988;16:83–91.
15. Borenstein DG, Lacks S, Wiesel SW. Cyclobenzaprine and naproxen versus naproxen alone in the treatment of acute low back pain and muscle spasm. *Clin Ther* 1990;12:125–31.
16. Yeung CK, Leung MC, Chow DH. The use of electro-acupuncture in conjunction with exercise for the treatment of chronic low-back pain. *J Altern Complement Med* 2003;9:479–90.
17. Furukawa T, Barbui C, Cipriani A, et al. Imputing missing standard deviations in meta-analyses can provide accurate results. *J Clin Epidemiol* 2006; 59:7–10.
18. Sutton AJ, Abrams KR, Jones DR, et al. *Methods for Meta-Analysis in Medical Research*. London: John Wiley; 2000:29–31.
19. Hedges LV, Olkin I. Nonparametric estimators of effect size in meta-analysis. *Psychol Bull* 1984;96:573–80.
20. Berkey CS, Hoaglin DC, Mosteller F, et al. A random-effects regression model for meta-analysis. *Stat Med* 1995;14:395–411.
21. *Stata Statistical Software Manual [computer program]*. College Station, TX: Stata Corp; 2006.
22. Furlan AD, van Tulder MW, Cherkin DC, et al. Acupuncture and dry-needling for low back pain. *Cochrane Database Syst Rev* 2005:CD001351.
23. Hayden JA, van Tulder MW, Malmivaara A, et al. Exercise therapy for treatment of non-specific low back pain. *Cochrane Database Syst Rev* 2005: CD000335.
24. Schonstein E, Kenny DT, Keating J, et al. Work conditioning, work hardening and functional restoration for workers with back and neck pain. *Cochrane Database Syst Rev* 2003:CD001822.

25. Assendelft WJ, Morton SC, Yu Emily I, et al. Spinal manipulative therapy for low-back pain. *Cochrane Database Syst Rev* 2004:CD000447.
26. Heymans MW, van Tulder MW, Esmail R, et al. Back schools for non-specific low-back pain. *Cochrane Database Syst Rev* 2004:CD000261.
27. Furlan AD, Brosseau L, Imamura M, et al. Massage for low-back pain. *Cochrane Database Syst Rev* 2002:CD001929.
28. Hagen KB, Hilde G, Jamtvedt G, et al. Bed rest for acute low-back pain and sciatica. *Cochrane Database Syst Rev* 2004:CD001254.
29. Ostelo RW, van Tulder MW, Vlaeyen JWS, et al. Behavioural treatment for chronic low-back pain. *Cochrane Database Syst Rev* 2005:CD002014.
30. van Tulder MW, Touray T, Furlan AD, et al. Muscle relaxants for non-specific low-back pain. *Cochrane Database Syst Rev* 2003:CD004252.
31. van Tulder MW, Blomberg SEI, de Vet HCW, et al. Traction for low-back pain with or without radiating symptoms (Protocol). *Cochrane Database Syst Rev* 2001.
32. Urrútia G, Burton AK, Morral A, et al. Neuroreflexotherapy for non-specific low-back pain. *Cochrane Database Syst Rev* 2004:CD003009.
33. Khadilkar A, Milne S, Brosseau L, et al. Transcutaneous electrical nerve stimulation (TENS) for chronic low-back pain. *Cochrane Database Syst Rev* 2005:CD003008.
34. Van Tulder MW, Scholten RJ, Koes BW, et al. Non-steroidal anti-inflammatory drugs for low back pain: a systematic review within the framework of the Cochrane Collaboration Back Review Group. *Spine* 2000;25:2501–13.
35. Hagen KB, Hilde G, Jamtvedt G, et al. The cochrane review of advice to stay active as a single treatment for low back pain and sciatica. *Spine* 2002;27: 1736–41.
36. Guzman J, Esmail R, Karjalainen K, et al. Multidisciplinary rehabilitation for chronic low back pain: systematic review. *BMJ* 2001;322:1511–6.
37. Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *BMJ* 2001;323:42–6.
38. Emerson JD, Burdick E, Hoaglin DC, et al. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Control Clin Trials* 1990;11:339–52.
39. Verhagen AP, de Vet HC, de Bie RA, et al. Balneotherapy and quality assessment: interobserver reliability of the Maastricht criteria list and the need for blinded quality assessment. *J Clin Epidemiol* 1998;51:335–41.